УДК 004

# ИНФОРМАЦИОННЫЕ СИСТЕМЫ ОБРАБОТКИ И СЖАТИЯ ТЕКСТА

**В.В. ЕФРЕМОВ**
**И.Н. ЕФРЕМОВА**
**В.В. СЕРЕБРОВСКИЙ**
**А.А. ЧЕРЕПАНОВ**

***Юго-Западный государственный университет***

***e-mail:***
***kafedra-ipm@mail.ru***
***Cha84@mail.ru***

Рассматривается использование продукционного направления обработки изображения для задач сжатия символьной информации и методику для оценки корректности таких систем продукций, повышающий безопасность вычислительных систем для исследуемой проблематики.

Ключевые слова: обработка символьной информации, сжатие информации, системы продукций.

В настоящее время методы технической обработки изображения, находят все большее применение в задачах народного хозяйства. Среди типичных задач обработки изображения особо стоит выделить распознавание текста. Распознавание текста широко используется для конвертации книг и документов в электронный вид, для автоматизации систем учёта в бизнесе или для публикации текста на веб-странице. Распознавание позволяет редактировать текст, осуществлять поиск слова или фразы, хранить его в более компактной форме, демонстрировать или распечатывать материал, не теряя качества, анализировать информацию, а также применять к тексту электронный перевод, форматирование или преобразование в речь.

На сегодняшний момент во многих задачах обработки символьной информации эффективно применяется продукционный подход [1]. Указанный подход может найти важное применение в проблеме сжатия символьной информации, которая занимает важное место среди задач обработки символов. В связи с этим, известные способы сопоставления с множеством эталонных объектов, например приведенных в источнике [2], при проверке соответствия условиям параллельных продукционных систем не учитывают особенности задачи сжатия [3]. Предлагается применять продукционную систему для сжатия, задавая аналогично множественному поисковому запросу – m образцов, и одновременно для адаптивного сжатия определяя m соответствующих им модификаторов. Такая методология повысит результативность сжатия символьной информации, например, при применении на продукционных машинах.

Анализ достоверности системы продукций для сжатия имеет свои особенности, связанные со следующими условиями корректности.

Система продукций обеспечивает полную модификацию.

Каждый символ входного слова модифицируется не более одного раза.

Допустим существует служебный алфавит A* , каждому символу которого приведено в соответствие последовательность символов(в искомую последовательность может входить даже один единственный символ), в обрабатываемом алфавите A, который состоит из анализируемых символов, можно вставить любое слово.

Определение. Систему продукций будем считать нормальной, если антецеденты однозначно соответствуют символам алфавита A*.

Теорема 1. Система продукций выполнит полную модификацию только тогда, когда она будет нормальной.

Доказательство. Допустим образцы не составляют нормальный класс и существует символ ξ отличный от пустого слова Λ, принадлежащий алфавиту A* и графически не равный ни одному образцу $S2^i$ , где i=(1..n), n — количество образцов.. Допустим слово $S1 = S2^1 \xi S2^2$. Тогда аннуляция вхождений образцов $S2^i \to \wedge$ приведет к слову $S1 = \xi$.

Модификация по определению будет полной тогда, когда $\xi = \wedge$ или существует образец графически равный ξ. Оба условия противоречат условиям теоремы. Теорема доказана.

Рассмотрим все возможные варианты пересечения образцов, приводящие к двукратному нахождению вхождения символов слова.

1. $S2^i = S2^j\ R2$ (вхождение слева),

2. $S2^i = R1\ S2^j$ (вхождение справа),

3. $S2^i = R1\ S2^j\ R2$ (вхождение по центру),

4. $S2^{ik} = S2^{jH}$ (пересечение), где R1, R2 – произвольные слова в алфавите A; $S2^{ik}, S2^{jH}$ - соответственно, конечный и начальный фрагмент слов $S2^i$ и $S2^j$.

Варианты 1, 3 и 4 вызывают неправильную модификацию при параллельном выполнении рассматриваемой системы продукций с ее спецификой в связи с тем, что графическое равенство фрагментов входного слова вхождению образцов друг в друга или их пересечению, приведет к двукратной модификации. Вариант 2 отличается тем, что конечные позиции вхождения образцов будут одинаковы. В этом случае для обеспечения корректности нужно ввести разные уровни приоритетов для образцов – высший {i} для более длинного $S2^i$, согласно которым срабатывает продукция с более высоким приоритетом {i}>{j}.

Теорема 2. Система продукций является неправильной для параллельного выполнения и однонаправленного потока данных, когда верно условие:

$$\begin{aligned}&\left(S2^i = S2^j R2\right)\vee\\ &\vee\left(S2^i = R1 S2^j R2\right)\vee\\ &\vee\left(S2^{ik} = S2^{JH}\right)\vee\\ &\vee\left[\left(S2^i = R1 S2^j\right)\&\left(\{i\}\le\{j\}\right)\right]=1\end{aligned}$$

Доказательство. Предположим, что верным является первый член дизъюнкции 1 и система является корректной, т.е. каждый символ модифицируется только один раз. Предположим, что слово $S1 = S2^j R2$. Для слова $S1$ будет детектировано вхождение образов $S2^j$ и $S2^i$ и фрагмент слова $S2^j$ причислен к обоим из них, т.е. будет модифицирован дважды, что противоречит условию корректности 2.

Предположим, что верным является второй член дизъюнкции 1 и система является корректной. Предположим, что слово $S1 = R1\ S2^j\ R2$. Для слова S1 будет детектировано вхождение образов $S2^j$ и $S2^i$, а участок слова $S2^j$ причислен к обоим из них, т.е. будет модифицирован дважды, что противоречит условию корректности 2.

Предположим, что $S2^i = R1\ \xi,\ S2^j = \xi\ R2$ (верным является третий член дизъюнкции 1) и система является корректной. Предположим, что слово $S1 = R1\ \xi\ R2$. Для слова S1 будет детектировано вхождение образов $S2^j$ и $S2^i$, а участок слова S2 соответствует обоим, т.е. будет модифицирован дважды, что противоречит условию корректности 2.

Предположим, что верным является четвертый член дизъюнкции 1 и система является корректной. Допустим слово $S1 = R1\ S2^j$. Для слова S1 будет зафиксировано вхождение образов $S2^j$ и $S2^i$ и фрагмент слова $S2^j$ соотнесен к обоим из них. Если {i}={j} фрагмент будет модифицирован дважды, что противоречит условию корректности 2. Если {i} < {j}, фрагмент слова R1 не будет модифицирован, что противоречит условию корректности 1. Теорема доказана.

Рассмотрение всех возможных вариантов пересечения образцов является основанием для следующего высказывания.

Следствие теоремы 2. Если условие 1 не выполняется, параллельная система продукций модифицирует каждый символ входного слова не более одного раза.

Таким образом, система продукций, удовлетворяющая условию теоремы 1 и условию следствия теоремы 2, является корректной для процедуры сжатия, а сами теоремы являются инструментальным базисом для проверки корректности систем продукций для сжатия символьной информации.

**Список литературы**

1. Довгаль В.М. Методы модификации формальных систем обработки символьной информации. Курск, 1996. 115 с.

2. Керекеша В.В. Ассоциативные устройства для реализации систем продукций: автореф. ... дис. канд. техн. наук. Курск, 1995.

3. Е.Г. Жиляков. Об эффективности алгоритма субполосного выделения контуров на изображении Е.Г. Жиляков, А.А. Черноморец, В.А. Голощапова, А.Н. Заливин // Научные ведомости Белгородского государственного университета № 15 (158) 2013, Выпуск 27/1. С.128-134.

## INFORMATION SYSTEMS OF PROCESSING AND TEXT COMPRESSION

**V.V. EFREMOV**
**I.N. EFREMOVA**
**V.V. SEREBROVSKY**
**A.A. CHEREPANOV**

***Southwest State University***

***e-mail:***
***kafedra-ipm@mail.ru***
***Cha84@mail.ru***

Use of the productional direction of processing of the image for problems of compression of symbolical information and a technique for an assessment of a correctness of such systems of the produktion, raising safety of computing systems on investigated perspectives is considered.

Key words: processing of symbolical information, compression of information, system of produktion.